EXEMPLAR-BASED NOISE ROBUST AUTOMATIC SPEECH RECOGNITION USING MODULATION SPECTROGRAM FEATURES

Deepak Baby[⋆]*, Tuomas Virtanen*[†]*, Jort F. Gemmeke*[⋆]*, Tom Barker*[†] *and Hugo Van hamme*[⋆]

[⋆]Department ESAT, KU Leuven, Belgium
[†]Department of Signal Processing, Tampere University of Technology, Finland
Deepak.Baby@esat.kuleuven.be

ABSTRACT

We propose a novel exemplar-based feature enhancement method for automatic speech recognition which uses coupled dictionaries: an input dictionary containing atoms sampled in the modulation (envelope) spectrogram domain and an output dictionary with atoms in the Mel or full-resolution frequency domain. The input modulation representation is chosen for its separation properties of speech and noise and for its relation with human auditory processing. The output representation is one which can be processed by the ASR back-end. The proposed method was investigated on the AURORA-2 and AURORA-4 databases and improved word error rates (WER) were obtained when compared to the system which uses Mel features in the input exemplars. The paper also proposes a hybrid system which combines the baseline and the proposed algorithm on the AURORA-2 database which in turn also yielded improvement over both the algorithms.

Index Terms: modulation envelope, coupled dictionaries, non-negative matrix factorisation, automatic speech recognition

1. INTRODUCTION

The paradigm of automatic speech recognition (ASR) has received a widespread interest with advancements in signal processing and automation in assistive devices for the handicapped, archive indexing, electronic gadgets, etc.. However, these systems are less robust to noisy conditions especially when the signal-to-noise ratio (SNR) falls towards 0dB or less. The enhancement of speech features by separating out the artefacts introduced by noise thus has become an essential part of noise-robust ASR. Here, we aim to address the feature enhancement of single channel speech data in the presence of non-stationary noise by using time-frequency models of speech and noise spanning hundreds of milli-seconds rather than working with shorter time contexts as used in methods like spectral subtraction [1], vector Taylor series approximation [2] etc.

Spectral factorisation methods based on non-negative matrix factorisation (NMF) have been successfully used for noise robust ASR [3–5], thanks to the ability of NMF to obtain sparse compositional model of features using the dictionary atoms or exemplars. NMF also claims to model human perception as non-negative data is quite naturally occurring in human auditory and visual processing [6]. Motivated from these two ideas and considering the fact that the performance of humans in recognizing noisy speech is far better than the current noise robust ASR systems, we propose using auditory models to extract the features and do the feature enhancement.

The author has done this work during his stay at Tampere University of Technology, Finland which was funded with support from the European Commission under Contract FP7-PEOPLE-2011-290000.

Conventional exemplar-based ASR systems operate on the magnitude or power spectrogram of the noisy speech to obtain features like Mel energies [7], Gabor filtering [8] etc., whereas humans rely on the low frequency amplitude modulation variation within frequency bands [9] which are computationally modelled as modulation envelope features. These modulation patterns play a key role in the higher level human auditory processing [10]. It is also shown that using the spectrogram of the modulation envelope called modulation spectrogram (MS) [11] yields a better representation because human speech contains modulations of very low frequency of the order of 20 Hz [12] and is found to be useful for blind source separation [13] and noise-robust ASR [14].

Traditional exemplar-based systems obtain a Wiener filter from the compositional model and use it for feature enhancement (FE). Earlier approaches used the Mel [3] or DFT (refers to the magnitude of the discrete-Fourier transform throughout this paper) [15] features to obtain the decomposition so that the Wiener filter can be directly obtained in a feature space that can then be used to find the Mel frequency cepstral coefficients (MFCCs) for the ASR back-end. Knowing that the speech and noise often have different modulation envelopes, a better separation between speech and noise is expected after applying NMF on MS features. However, obtaining the modulation spectrogram involves non-linear operations [11] which makes it hard to invert to the time domain. The idea of using coupled dictionaries has been successfully used to overcome this drawback for speech enhancement with MS features [16]. Here it is used to obtain a reliable reconstruction of the underlying Mel/DFT features which can then be used for MFCC based ASR.

The proposed algorithm is compared with a Mel-based exemplar system on the AURORA-2 [7] and the AURORA-4 [17] databases and gives better performance in terms of word error rate (WER). For the AURORA-2 database, a hybrid system is also introduced in this paper by combining the baseline and the proposed algorithms to exploit the complementarity of the recognition results which also yielded performance improvement over both systems. To our knowledge, this is the first use of modulation features in exemplar-based noise robust ASR, and the WERs obtained for the hybrid system on the AURORA-2 database are in fact among the best results ever reported on the database with and without multi-condition training (reported in [7]).

2. FEATURE ENHANCEMENT USING NMF

2.1. NMF based compositional model for noisy speech

Compositional models based on NMF aim to represent the features extracted from the noisy speech as a sparse non-negative linear combination of speech and noise atoms or exemplars. The exemplars

may span time windows of length T_t and then be reshaped to a long vector to capture the temporal variation. We denote $\mathbf{A}_s$ and $\mathbf{A}_n$ as the dictionary matrices, the columns of which contain the exemplars corresponding to speech and noise respectively. The noisy utterance is also converted to a similar representation of features by means of a sliding window approach over the length of the utterance with window of duration T_t and hop size T_h. The features belonging to each window are then reshaped to a vector and are stored as the columns of a matrix [3], $\mathbf{\Psi}$ which is then approximated as:

$$\mathbf{\Psi} \approx \begin{bmatrix}\mathbf{A}_s & \mathbf{A}_n\end{bmatrix} \begin{bmatrix}\mathbf{X}_s \\ \mathbf{X}_n\end{bmatrix} = \mathbf{AX} \quad s.t. \quad \mathbf{X} \geq 0. \tag{1}$$

The approximation is done using NMF which minimizes the Kullback-Leibler divergence between $\mathbf{\Psi}$ and $\mathbf{AX}$ with additional sparsity constraint on $\mathbf{X}$ as in [18].

2.2. Baseline system using Mel exemplars

The exemplar-based ASR system explained in [18], which uses Mel integrated magnitude spectra of acoustical data as exemplars stored in the *Mel dictionary*, $\mathbf{A}^{\text{mel}} = \begin{bmatrix}\mathbf{A}_s^{\text{mel}} & \mathbf{A}_n^{\text{mel}}\end{bmatrix}$, is considered the baseline system in our experiments. Let $\mathbf{M}$ be the DFT-to-Mel matrix which contains the magnitude response of B Mel bands along its rows.

For testing, the compositional model for the noisy data in the Mel exemplar space is obtained using $\mathbf{A}^{\text{mel}}$ which yields the corresponding activations $\mathbf{X}_s^{\text{mel}}$ and $\mathbf{X}_n^{\text{mel}}$. The windowed estimates for speech and noise are then obtained as $\hat{\mathbf{s}} = \mathbf{A}_c^{\text{mel}}\mathbf{X}_c^{\text{mel}}$ and $\hat{\mathbf{n}} = \mathbf{A}_n^{\text{mel}}\mathbf{X}_n^{\text{mel}}$ respectively in the Mel exemplar space. The frame-level Mel estimates, $[\hat{\mathbf{s}}]^\star$ and $[\hat{\mathbf{n}}]^\star$ are then found after removing the windowing effect by averaging estimates belonging to overlapping windows [18]. The resulting Wiener filter is then generated, by element-wise division, as

$$\mathbf{W}^{\text{mel}} = [\hat{\mathbf{s}}]^\star \oslash [\hat{\mathbf{s}} + \hat{\mathbf{n}}]^\star, \tag{2}$$

called the *Mel filter*. The noisy Mel features are enhanced using $\mathbf{W}$ and are then fed to the ASR system to obtain the baseline results.

To obtain the ASR results on the AURORA-4 database, the "recipe" HMM-GMM recognizer in the Kaldi toolkit [19] is used. For the baseline system, the frame-level Mel estimates are used to estimate the Wiener filter in the DFT space using the pseudo-inverse of the DFT-to-Mel matrix, $\mathbf{M}^\dagger$, as

$$\mathbf{W}^{\text{dft}} = \left(\mathbf{M}^\dagger [\hat{\mathbf{s}}]^\star\right) \oslash \left(\mathbf{M}^\dagger [\hat{\mathbf{s}} + \hat{\mathbf{n}}]^\star\right), \tag{3}$$

called the *DFT filter*. This is then used for speech enhancement on the AURORA-4 data which are then fed to the Kaldi recognizer to obtain the baseline results [17].

3. PROPOSED METHOD USING MS FEATURES

The proposed method based on the modulation spectrogram features using the coupled dictionaries along with the design requirements are discussed in this section. The algorithm aims at finding a Wiener filter to enhance the noisy Mel/DFT features utilising the signal separation capabilities of modulation spectrogram features using NMF.

3.1. Modulation spectrogram features

The MS representation was proposed as part of a computational model for human hearing which relies on the low frequency amplitude modulations within various frequency bands [9] which are called modulation envelopes. Let B be the number of frequency bands considered. The MS representation for acoustical data is obtained by taking the STFT of the modulation envelopes corresponding to each frequency band [11]. The STFT is obtained with a window length and hop size of t_w^{MS} and t_h^{MS} seconds respectively. For non-negativity, only the magnitude of the STFT is considered.

Because of the low-pass filtering operation, only very few lower bins of the MS will contain significant energy and it is possible to truncate each of the MS to the lowest b bins [13]. All these truncated MS of size $b \times T$ each are then stacked to obtain a matrix of size $(B \cdot b) \times T$ [16] which are referred to as *MS features*.

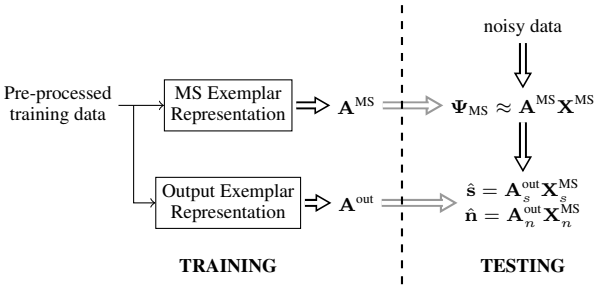

Fig. 1. *Block digram overview of the proposed system using modulation spectrogram features and coupled dictionaries.*

3.2. Method

In the proposed system, the NMF-based decomposition is obtained in the MS feature space which serves as the front-end of the ASR system. However, as mentioned before, even though the modulation features are known to yield different patterns for speech and noise, its utility in this framework is limited because the phase information is discarded to ensure non-negativity which results in the non-uniqueness of an inverse operation to the feature space that is used by ASR systems (eg. MFCCs). The proposed method using coupled dictionaries overcomes this drawback and yields a direct and reliable reconstruction of the underlying Mel/DFT output features which is summarised in Figure 1, where $\mathbf{\Psi}_{\text{MS}}$ for the noisy speech represented in the MS exemplar domain. All matrix divisions should be considered element-wise.

To obtain the dictionaries, each of the coupled exemplars for the MS and the Mel/DFT dictionaries are extracted from the same piece of training data which spans multiple frames of length T, followed by reshaping to form a vector. This will result in speech and noise dictionaries each for the MS and Mel/DFT output exemplar representations which are denoted as $\mathbf{A}_s^{\text{MS}}$, $\mathbf{A}_n^{\text{MS}}$, $\mathbf{A}_s^{\text{out}}$ and $\mathbf{A}_n^{\text{out}}$ respectively.

During the test phase, the NMF-based decomposition in the MS exemplar space yields the activations $\mathbf{X}^{\text{MS}}$. These activations are then applied to the coupled output dictionaries to reconstruct the underlying windowed Mel/DFT estimates corresponding to speech and noise. The corresponding Wiener filter is then obtained as given in Section 2.2. Because we are finding the Wiener filter which captures the relative energies between speech and noise, reusing the activations from the coupled feature space is a valid choice as long as the mapping is one-to-one. But in this framework, since the MS features lack phase information, there can be multiple Mel/DFT output exemplars that can lead to the same MS exemplar making the mapping many-to-one. However, this issue can be addressed by adjusting the parameters used to obtain the MS features and make the mapping nearly one-to-one which will be discussed next.

3.3. Design requirements

Notice that the STFT operations in the MS and Output layers can use different window lengths and hop sizes which adds to the flexibility of the framework. This allows the designer to use baseline settings of Section 2.2 for finding the Mel/DFT data and tune the parameters in the MS domain to obtain the optimum separation. In this paper, we used the baseline settings for finding the Mel features for a reliable comparison. But the window length, t_w^{MS} and hop size, t_h^{MS} for finding MS features should be tuned to address the many-to-one mapping between the coupled MS and output exemplars.

Choice of window length In the procedure to obtain the MS features, there is a low-pass filtering operation which typically has a 3dB cut-off frequency of around 20-40 Hz based on the knowledge of the human auditory system [12] similar to RASTA filtering [20]. The window length determines the number of modulation frequency bands b that are located below this cut-off frequency. Long windows lead to many bands which yield an accurate representation of the modulation analysis, but also produce a risk of overfitting to the modulation properties of the training noise types. So a compromise must be pursued.

Temporal oversampling Any circular temporal shift of the DFT spectrogram will lead to the same magnitude modulation spectrogram (modulo the window length). However, as pointed out in [21], given an oversampled sequence of spectra obtained from overlapping windows, it is possible to reduce this ambiguity greatly. Therefore, the MS dictionary atoms store magnitude modulation spectra of successive overlapping windows sampled at a rate high enough to reduce the ambiguity of the MS to Mel/DFT mapping. Thus, a temporal oversampling is done to obtain the MS features using small hop sizes with long window lengths to realise one-to-one mapping.

3.4. Hybrid system

In the proposed approach and the baseline algorithm, the activations that lead to the two Wiener filters are derived from two different feature spaces and both the algorithms were found to give different results. There exist several ways to combine results from two systems like assuming independence and then balance the two streams [7], minimum error based approach [22], etc. To avoid extra parameters, we propose to combine the algorithms by simply multiplying the likelihoods, obtained using the enhanced Mel features and the trained acoustic model, of the two streams in a HMM based recognizer and do the decoding. i.e.,

$$p''(y_t|q_t) = p(y_t|q_t)p'(y_t|q_t) \tag{4}$$

where, p and p' are the likelihoods for the observation y_t given the HMM state q_t resulting from the Mel and MS streams respectively.

4. EVALUATION EXPERIMENTS

4.1. AURORA-2 database

In this work, we used test sets A and B of the AURORA-2 database. The database contains utterances of digits from '0-9' and 'oh' with no pre-defined grammar. The training data contains clean speech utterances and noisy data corrupted with additive noise of four noise types (subway, babble, car and exhibition hall). Test set A contains 1 clean subset and 6 subsets of each of the four noise types in the training set with varying SNRs ($-5, 0, 5, 10, 15, 20$ in dB), resulting in a total of 24 noisy subsets. Test set B also contains the same number of subsets but with four different additive noise types (restaurant, train station, street and airport). The WERs obtained after taking the average over the four noise types for clean speech, -5dB and combined average of SNRs ranging from 20-0 dB are presented.

The ASR back-end used a GMM-HMM-based recognizer using MFCC features. The HMM topology had 16 states describing each digit and 3 states for silence leading to a total of 179 states. The GMM model was trained on MFCCs with 13 static features along with their delta and delta-delta time differences resulting in a 39 dimensional feature space. The emission probabilities of each of the HMM state were modelled using a GMM of 32 Gaussians with diagonal covariance. For the viterbi decoder, an HMM topology where all the words have the same word entrance penalties, which was doubled for the hybrid approach, was used.

4.2. AURORA-4 database

AURORA-4 is a large-vocabulary continuous speech database based on the WSJ-0 corpus of read speech. In this work, we used the single microphone case with clean and six additive noise testing conditions. The clean speech set (test 01) contains 330 utterances with noisy test set containing its six noisy versions (test 02 - test 07) added with car, babble, restaurant, airport, street and train noises added artificially at varying SNRs between 5 and 15 dB. For training the acoustic model, 7138 utterances each of clean speech (clean training set) and multi-noise training (MNT) set containing both clean and noisy data are used.

The ASR results are obtained using the recipe recognizer in the Kaldi toolkit which uses a HMM-GMM system based on context-dependent tied-state triphone models. The setting has around 2000 distinct HMM states with three HMM states per model. For feature decorrelation, maximum-likelihood linear transform (MLLT) [23] and linear discriminant analysis (LDA) [24] were applied on stacked MFCC feature vectors (13 coefficients over 7 consecutive frames), reducing the 91-dimensional vetor to 40 dimensions.

4.3. NMF based enhancement

The coupled dictionaries were created using the procedure explained in Section 3 with the parameters used in [18]. A temporal context T_t of 300 ms and 150 ms was used to obtain the exemplars for the AURORA-2 and AURORA-4 databases, respectively. The noise data for training is extracted from the noisy training set using the procedure described in [18]. Then for every clean and noise data in the training set, five random pieces of T_t ms data were taken and obtained the corresponding Mel, DFT and MS exemplars. No supervision was done to avoid the overlap between the chosen random pieces of data or to remove silence.

The training data was pre-processed by removing the dc component and applying a pre-emphasis filter (of coefficient 0.97), before extracting the exemplars. To obtain the Mel exemplars, the magnitude spectrogram of the pre-processed training data was obtained using $t_w^{\text{mel}} = 25$ ms and $t_h^{\text{mel}} = 10$ ms. The Mel features were then obtained for each frame after Mel integration with B channels. The number of Mel bands used were 23 and 40 for the AURORA-2 and AURORA-4 databases, respectively.

To create the MS exemplars, equivalent rectangular bandwidth filter banks (B channels) implemented using Slaney's toolbox [25] were applied on the pre-processed data. The filter-bank outputs were then half-wave rectified and low-pass filtered at a 3dB cut-off frequency of choice (20 to 40 Hz). Then for each of the modulation envelopes, the magnitude spectrogram was obtained for different t_w^{MS} keeping $t_h^{\text{MS}} = t_h^{\text{mel}} = 10$ ms. Each of the spectrograms was then

Experiments	clean	*test set A* (20-0)	-5	*test set B* (20-0)	-5
Baseline	0.2	5.3	31.0	**8.0**	**55.3**
MS Experiments					
64ms;20Hz;4bins	**0.0**	4.4	**30.5**	8.2	62.7
64ms;30Hz;5bins	**0.0**	4.2	30.5	8.2	63.1
64ms;40Hz;5bins	**0.0**	4.4	30.8	8.1	62.9
128ms;30Hz;8bins	**0.0**	**4.1**	31.3	8.4	69.3

Table 1. *WER in % obtained with GMM trained on clean training data as a function of SNR in dB on a subset of 100 files in the AURORA-2 database. The results obtained for various window lengths and cut-off frequencies with appropriate choice for the number of bins, shown in order in the first column, are given.*

truncated to b bins of appropriate length ranging from 4-6. These are then stacked as explained in Section 3.1 which is then reshaped to get the MS exemplar.

For obtaining the compositional model on AURORA-2, NMF with sparsity constraint was used with 10000 exemplars each for the speech and noise dictionary. A sparsity penalty of 1.5 and 0.5 are used for the Mel speech and noise exemplars respectively as given in [18]. For the NMF with MS exemplars, best results were obtained for speech and noise exemplar sparsity penalties of 1.75 and 0.75 respectively, after a grid-search in the range $[0, 3]$ on a subset of 100 files in the test set (development set). The Mel exemplar space is used to create the coupled output dictionary and the enhanced Mel features were fed to the ASR back-end.

In the AURORA-4 setting, 10000 speech and 5000 noise exemplars were used to create the dictionary. In contrast to the AURORA-2 setting, the noise sparsity was set at 0.5 times the speech sparsity as in [17]. The Mel setting used a speech sparsity penalty of 0.075 times the average L_1 norm of the dictionary, whereas the MS setting used a ratio of 0.05. As mentioned before, experiments on AURORA-4 were evaluated by first obtaining the enhanced speech data and then feeding it to the Kaldi toolkit. For the baseline results, the DFT Wiener filter is obtained as in Section 2.2. For the proposed method, the coupled DFT output dictionary was used to directly obtain the DFT Wiener filter following the decomposition in the MS space. An additional experiment is also conducted, where the decomposition is done on the Mel space followed by the direct reconstruction of the DFT filter using its coupled output DFT dictionary, to show the effectiveness of coupled dictionaries in bypassing the artefacts introduced by the pseudo-inverse step.

5. RESULTS AND DISCUSSION

5.1. Analysis of the MS system on AURORA-2

To reduce the experimentation time, pilot experiments were conducted on a subset of 100 files, which is disjoint with the development set, to find the best parameters for the proposed system. An acoustic model trained on the clean training set was used. The baseline results were obtained using the Mel filter $\mathbf{W}$ described in Section 2.2 with 700 multiplicative update iterations for the NMF. The results for the MS sytem were obtained for varying t_w^{MS} and low-pass cut-off frequency which are tabulated in Table 1. It can be seen that the proposed system yields improved performance for matched noise types and performs poorer for unseen cases. This can be attributed to the increased dimensionality of the MS features which leads to very accurate modelling of matched noise cases. For SNR -5dB, the MS features perform as good as the baseline for test set A and as the SNR increases, MS features result in better separation and impressive WER improvements. It is also observed that varying the low-pass cut-off frequency does not have much effect on the performance.

Experiments	clean	*test set A* (20-0)	-5	*test set B* (20-0)	-5
Baseline	0.5	3.1	25.1	6.4	**52.8**
MS System	**0.4**	**2.4**	21.1	7.5	62.4
Hybrid	**0.4**	**2.4**	**20.6**	**6.1**	54.2

Table 2. *WER in % obtained with GMM trained on multi-condition training set as a function of SNR in dB on the AURORA-2 database.*

For test set B, the WER for SNR -5dB is far above that of the baseline but improves as SNR increases. It is observed that if we take the average over SNRs $5, 10, 15$ and 20dB, the WER is better than that of the baseline system. i.e., the MS features perform well for the unseen cases also if the SNR is sufficiently high. Increasing the hop size yielded an average WER of 40% for test set A, SNR -5dB (not shown) which was expected because of the poorer mapping between the two domains.

It is also observed that, as we increase the window length to 128 ms and take 8 bins per channel, the performance improves for seen noise cases because the signal model is even more accurate, and because of the same reason fails for unseen noise types as expected. Increasing the window length and number of bins used can improve the performance for test set A, but suffers from two issues : poorer performance for unseen cases and larger dimensionality demands more exemplars in the dictionary which increases the complexity.

5.2. Hybrid MS-Mel results with multi-condition training

From the MS experiments, the case with 30 Hz cut-off, 64 ms window length and 5 bins per channel was chosen for obtaining the results on the complete test set. The acoustic model for the MS and Mel streams were then trained separately on MFCCs obtained after applying the MS and Mel feature enhancement respectively, on the multi-condition training set (referred to as *retraining*). The set of results thus obtained are given in Table 2. Notice that the baseline results are better than reported in [7] because of the larger dictionaries used.

It can be seen that both the proposed systems yield superior performance improvements over the baseline system for test set A. For the hybrid approach, there is WER improvement for unseen cases also, as the baseline likelihoods correct the errors introduced due to poor modelling of unseen noise in the MS domain. The results obtained for test set A with this approach are among the best results ever reported on the AURORA-2 database to date.

5.3. Results on AURORA-4 database

For the MS setting, a window length of 64 ms with 10 ms hop size was used to obtain the modulation spectrogram. These are then truncated to the lowest 5 bins to obtain the MS exemplars. The results obtained on the AURORA-4 database for various choices of input and output exemplar representations are tabulated in Table 3. The do nothing baseline results on the database are shown on the first row. The notation Mel† in the second row denotes the baseline system in [17] which uses pseudo-inverse of the DFT-to-Mel matrix to map the Mel estimates to the DFT space. The last two rows shows the proposed methods using coupled output DFT dictionaries.

$\mathbf{A}^{\text{in}}$	$\mathbf{A}^{\text{out}}$	Test Sets 01	02	03	04	05	06	07	Avg.
	GMM training using clean data								
-	-	**4.9**	19.4	36.8	46.0	45.1	33.0	43.6	32.7
Mel	Mel†	5.3	10.7	19.5	26.1	28.8	16.8	27.5	19.2
Mel	DFT	5.2	**7.7**	14.4	20.9	**15.8**	12.5	**15.2**	13.1
MS	DFT	5.1	9.8	**13.0**	**17.7**	16.7	**11.2**	16.1	**12.8**
	Training on MNT data (Not Retrained)								
-	-	6.6	7.3	12.6	24.1	17.4	11.6	16.9	13.8
Mel	Mel†	7.1	7.2	10.2	14.4	14.4	9.5	14.6	11.1
Mel	DFT	6.0	**6.3**	9.5	**11.6**	9.5	8.5	**9.1**	**8.6**
MS	DFT	**5.8**	6.9	**8.2**	**11.6**	**9.4**	**8.2**	10.1	**8.6**
	Training on MNT data (Retrained)								
-	-	6.6	7.3	12.6	24.1	17.4	11.6	16.9	13.8
Mel	Mel†	6.2	6.5	9.7	16.9	14.6	9.8	14.7	12.0
Mel	DFT	5.6	**5.6**	7.9	**10.9**	10.0	7.6	10.7	8.8
MS	DFT	**5.3**	5.9	**7.5**	11.0	**9.2**	**7.1**	**10.1**	**8.5**

Table 3. *WER in % obtained for various systems with GMM trained on clean, unenhanced multi-noise training data (MNT) and enhanced MNT data (Retrained) on the AURORA-4 database. Best scores are highlighted in bold font.*

It can be seen that, for all the cases the decomposition using the MS features followed by direct DFT reconstruction using the coupled dictionary yields better results. Similar to the observations made in the AURORA-2 database, these improvements can be attributed to the better speech and noise separation capability of the MS feature representation.

It was already observed in [17] that feeding NMF enhanced data can improve the ASR results. Here, it can be seen that the proposed Mel-DFT system is far better than the Mel-Mel† baseline system even though the decomposition in both systems is done in the Mel space. This work thus reveals that even though Mel features can lead to a good speech and noise separation, the use of pseudo-inverse to map the Mel features to the DFT space results in undesired artefacts and using coupled DFT dictionary can overcome this drawback.

6. CONCLUSIONS AND FUTURE WORK

This paper proposed two novel methods in the class of exemplar-based noise robust ASR algorithms. First, modulation features were introduced in the exemplar-based systems. Second, a coupled dictionary training was proposed, the use of which is not restricted to just the modulation domain. We also introduced a way to tackle the poor mapping between the two coupled domains by means of temporal oversampling. The use of a hybrid system was also investigated in this paper which on AURORA-2 database gave an average WERs of 2.4 % (0-20dB) and 20.6 % at -5 dB SNR on test set A and 6.1 % (0-20dB) on test set B, which are among the best results ever obtained on the database. The proposed technique was also investigated on the AURORA-4 database which reaffirmed the effectiveness of the coupled dictionary approach and modulation spectrogram features yielding highly significant WER improvements ($p < 0.001$).

This work opens a new line of research, to investigate the use of coupled dictionaries for other purposes also, for eg. signal enhancement. The effectiveness of using the enhanced data on the deep-neural network (DNN) based ASR setting also has to be investigated.

7. REFERENCES

[1] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 27, no. 2, pp. 113–120, April 1979.

[2] P. Moreno, B. Raj, and R. Stern, "A vector taylor series approach for environment-independent speech recognition," in *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, vol. 2, May 1996, pp. 733–736 vol. 2.

[3] J. F. Gemmeke and T. Virtanen, "Noise robust exemplar-based connected digit recognition," in *Proc. ICASSP*, March 2010, pp. 4546 –4549.

[4] F. Weninger, J. Geiger, M. Wöllmer, B. Schuller, and G. Rigoll, "The Munich 2011 CHiME challenge contribution: NMF-BLSTM speech enhancement and recognition for reverberated multisource environments," in *CHiME 2011 Workshop on Machine Listening in Multisource Environments*, September 2011.

[5] E. Yılmaz, J. F. Gemmeke, D. Van Compernolle, and H. Van hamme, "Noise-robust digit recognition with exemplar-based sparse representations of variable length," in *IEEE Workshop on Machine Learning for Signal Processing (MLSP)*, Santander, Spain, Sept. 2012.

[6] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, pp. 788–791, October 1999.

[7] J. F. Gemmeke and H. Van hamme, "Advances in noise robust digit recognition using hybrid exemplar based systems," in *Proc. INTERSPEECH*, 2012.

[8] A. Hurmalainen and T. Virtanen, "Modelling spectro-temporal dynamics in factorisation-based noise-robust automatic speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, 2012, pp. 4113–4116.

[9] C. Plack, *The sense of hearing*. Lawrence Erlbaum Associates Publishers, 2005.

[10] A.S. Bregman, *Auditory scene analysis: The perceptual organization of sound*. The MIT Press, 1990.

[11] S. Greenberg and B. Kingsbury, "The modulation spectrogram: in pursuit of an invariant representation of speech," in *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*, vol. 3, 1997, pp. 1647–1650 vol.3.

[12] C. E. Schreiner and J. V. Urbas, "Representation of amplitude modulation in the auditory cortex of the cat. I. The anterior auditory field (AAF)," *Hearing Research*, vol. 21, no. 3, pp. 227 – 241, 1986.

[13] T. Barker and T. Virtanen, "Non-negative tensor factorization of modulation spectrograms for monaural sound source separation," in *Proc. INTERSPEECH*, 2013.

[14] B. Kingsbury, N. Morgan, and S. Greenberg, "Robust speech recognition using the modulation spectrogram." *Speech Communication*, vol. 25, no. 1-3, pp. 117–132, 1998.

[15] T. Virtanen, R. Singh, and B. Raj, Eds., *Techniques for Noise Robustness in Automatic Speech Recognition*. Wiley, 2012.

[16] D. Baby, T. Virtanen, T. Barker, and H. Van hamme, "Coupled dictionary training for exemplar-based speech enhancement," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, 2014.

[17] J. T. Geiger, J. F. Gemmeke, B. Schuller, and G. Rigoll, "Investigating NMF speech enhancement for neural network based acoustic models," in *INTERSPEECH*. ISCA, 2014.

[18] J. F. Gemmeke, T. Virtanen, and A. Hurmalainen, "Exemplar-based sparse representations for noise robust automatic speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2067 –2080, Sept. 2011.

[19] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, Dec. 2011, iEEE Catalog No.: CFP11SRW-USB.

[20] H. Hermansky and N. Morgan, "Rasta processing of speech," *Speech and Audio Processing, IEEE Transactions on*, vol. 2, no. 4, pp. 578–589, Oct 1994.

[21] D. Griffin and J. Lim, "Signal estimation from modified short-time Fourier transform," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '83.*, vol. 8, 1983, pp. 804–807.

[22] J. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer Output Voting Error Reduction (ROVER)," in *Automatic Speech Recognition and Understanding, 1997. Proceedings., 1997 IEEE Workshop on*, 1997, pp. 347–354.

[23] G. Saon, M. Padmanabhan, R. Gopinath, and S. Chen, "Maximum likelihood discriminant feature spaces," in *Acoustics, Speech, and Signal Processing, 2000. ICASSP '00. Proceedings. 2000 IEEE International Conference on*, vol. 2, 2000, pp. II1129–II1132 vol.2.

[24] R. Haeb-Umbach and H. Ney, "Linear discriminant analysis for improved large vocabulary continuous speech recognition," in *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, vol. 1, Mar 1992, pp. 13–16 vol.1.

[25] M. Slaney, "Auditory toolbox version 2," *Interval Research Corporation*, vol. 10, p. 1998, 1998.